Supplementary Information - Early warnings of hazardous thunderstorms over Lake Victoria

Wim Thiery[1,2], Lukas Gudmundsson[1], Kristopher Bedka[3], Fredrick H.M. Semazzi[4], Stef Lhermitte[5], Patrick Willems[2,6], Nicole P.M. van Lipzig[6] and Sonia I Seneviratne[1]

[1] Institute for Atmospheric and Climate Science, ETH Zurich, Universitaetstrasse 16, 8092 Zurich, Switzerland
[2] Department of Hydrology and Hydraulic Engineering, Vrije Universiteit Brussels, Pleinlaan 2, 1050 Brussels, Belgium
[3] NASA Langley Research Center, 21 Langley Boulevard, Hampton, Virginia 23681, USA
[4] Department of Marine, Earth and Atmospheric Sciences, North Carolina State University, Raleigh, NC 27695, USA
[5] Department of Geoscience and Remote Sensing, Delft University of Technology, Stevinweg 1, 2600 GA Delft, The Netherlands.
[6] Department of Earth and Environmental Sciences, KU Leuven, Celestijnenlaan 200E, 3001 Leuven, Belgium.

E-mail: wim.thiery@env.ethz.ch

March 2017

Submitted to: *Environ. Res. Lett.*

The Odds Ratio (OR) is given by:

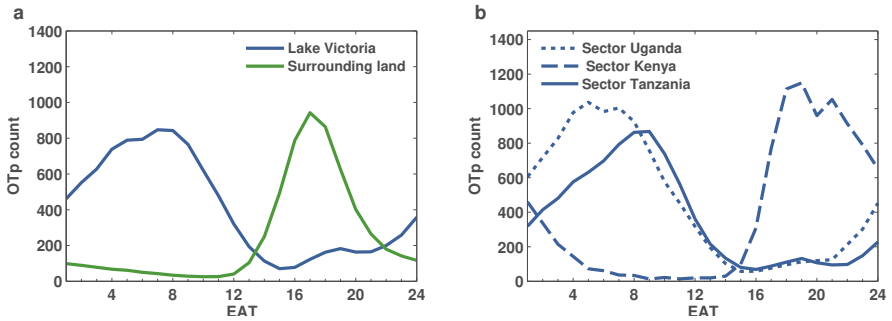
$$OR = \frac{TruePositives * TrueNegatives}{FalsePositives * FalseNegatives} \tag{1}$$

The Positive Predictive Value (PPV) is given by:

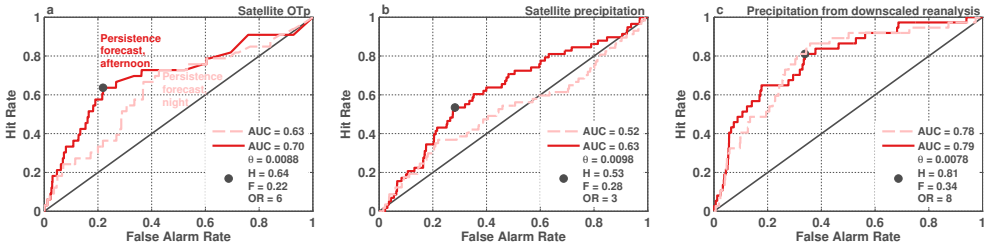
$$PPV = \frac{TruePositives}{TruePositives + FalsePositives} \tag{2}$$

The Negative Predictive Value (NPV) is given by:

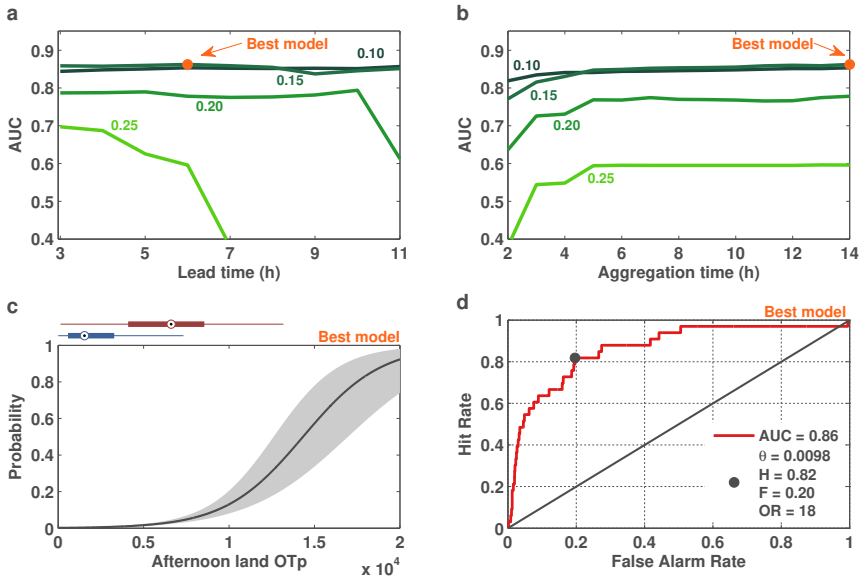
$$NPV = \frac{TrueNegatives}{TrueNegatives + FalseNegatives} \tag{3}$$



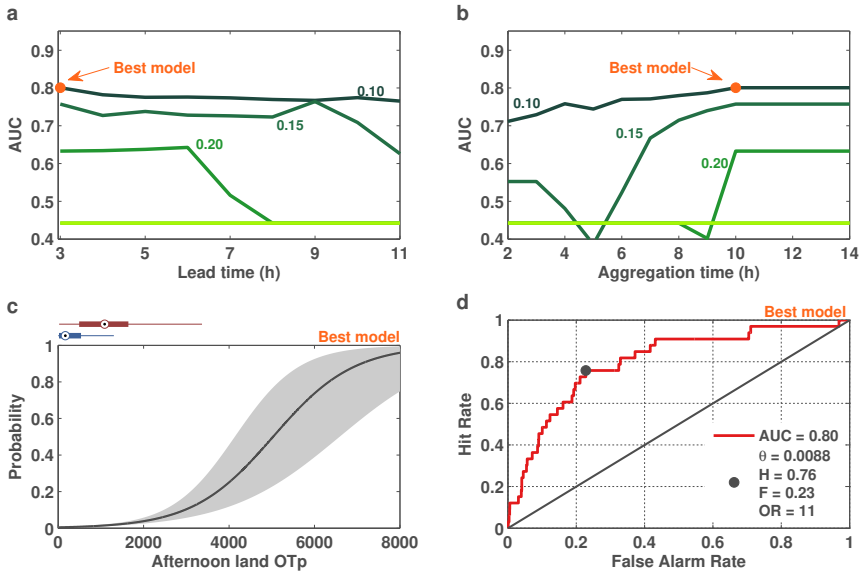
Supplementary Figure 1 | Diurnal storm cycle over Lake Victoria. (*a*) Hourly mean number of OTp detected over Lake Victoria and surrounding land (see figure 1c) for the period 2005-2013. (*b*) Same as *a*, but for the sectors of the lake belonging to the different countries.



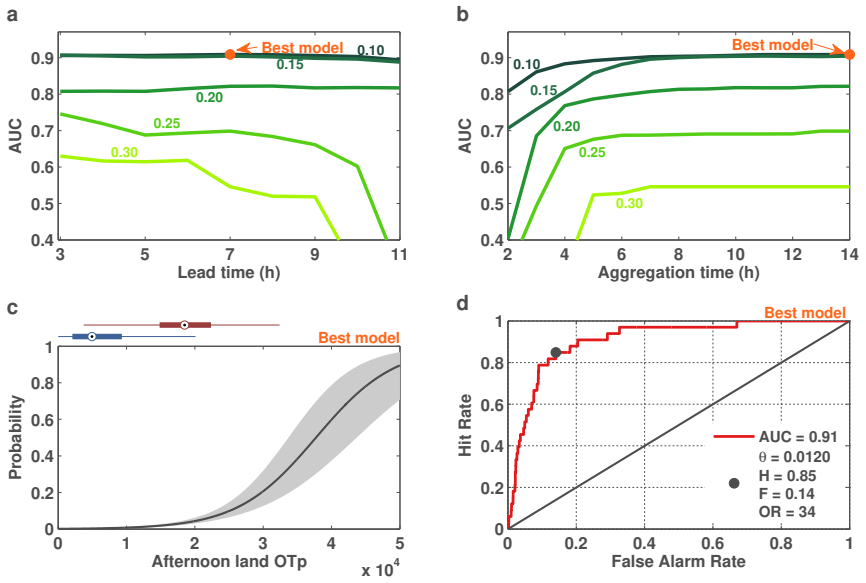
Supplementary Figure 2 | Limited value of persistence forecasts. Receiver operating characteristics (ROC) curves showing the trade-off between hit-rate and false-alarm rate for the predictions with difference threshold probabilities of the logistic regression model applied to (*a*) Satellite-based OT detections during 2005-2013, (*b*) Satellite-based precipitation observations during 1998-2013 from the Tropical Rainfall Measurement Mission (TRMM; 3B42 product) and (*c*) Precipitation during 1999-2008 as downscaled from ERA-interim by the regional climate model COSMO-CLM^2. Solid red lines indicate ROC curves using **afternoon lake** OTp/precipitation as predictor (initial configuration, see section 2.2), whereas dashed light red lines denote ROC curves using **nighttime lake** OTp/precipitation from the previous night as predictor. The ROC curve is estimated using leave-one-year-out-cross validation. Curves above the 1:1 line indicate that forecasts outperform random guessing. The black dots mark the location at which the difference between the hit rate and the false alarm rate is maximal. θ marks the threshold probability that was chosen to maximise this difference, H the hit rate, F the false alarm rate and OR the odds ratio at this point. The Area-Under-the-Curve AUC represents the dimensionless area under the ROC curve and is a measure for the overall skill of the regression model.



Supplementary Figure 3 | Optimizing the prediction system for Sector Uganda. (*a*) Area-Under-the-Curve *AUC* versus forecast lead time for various threshold correlations (aggregation time = 14h). (*b*) Area-Under-the-Curve *AUC* versus forecast aggregation time for various threshold correlations (lead time = 4h). Rank correlation thresholds are indicated for each line. Orange dots in (*a*) and (*b*) indicate the best model, i.e. the model with the highest AUC. (*c*) Logistic regression with optimised lead time, aggregation time and rank correlation threshold (see Table 1), modelling the probability of extreme nighttime OT detections on Lake Victoria as a function of afternoon OTp over selected land pixels (line: best estimate, shaded area: 95% confidence interval). Boxplots indicate the distribution (median, interquartile range) of the afternoon land OTp associated with non-extreme (blue) and extreme (red) nights. Whiskers extend to the last value located within a distance of 1.5 times the interquartile range from the 25th and 75th quantile, respectively. (*d*) Receiver operating characteristics (ROC) curve associated with the optimised logistic regression model shown in *c*.



Supplementary Figure 4 | Optimizing the prediction system for Sector Kenya. Same as Supplementary Figure **??**, but for Sector Kenya.



Supplementary Figure 5 | Optimizing the prediction system for Sector Tanzania. Same as Supplementary Figure **??**, but for Sector Tanzania.